

# Entropy monotonic spectral method for Boltzmann equation

Zhenning Cai\*, Yuwei Fan†, Lexing Ying‡

April 26, 2017

## Abstract

We propose a spectral method that discretizes the Boltzmann collision operator and satisfies a discrete version of the H-theorem. The method is obtained by modifying the existing Fourier spectral method to match a classical form of the discrete velocity method. It preserves the positivity of the solution on the Fourier collocation points and as a result satisfies the H-theorem. The fast algorithms appeared previously in the literature can be readily applied to this method to speed up the computation. A second-order convergence rate is validated by numerical experiments.

**Keywords:** Fourier spectral method; discrete velocity method; Boltzmann equation; modified Jackson filter; H-theorem; positivity.

## 1 Introduction

Gas kinetic theory describes the statistical behavior of a large number of gas molecules in the joint spatial and velocity space. It has been widely used to describe gases outside the hydrodynamic regime, for example in the field of rarefied gas dynamics. Based on the molecular chaos assumption, the Boltzmann equation

$$\frac{\partial F}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} F = \mathcal{Q}(F, F), \quad (1.1)$$

was derived from such a theory in [4] and turned out to be an accurate model in the gas dynamics. Here  $F(t, \mathbf{x}, \mathbf{v})$  is the mass density distribution function of the particles, depending on the time  $t \in \mathbb{R}^+$ , position  $\mathbf{x} \in \mathbb{R}^D$  ( $D \geq 2$ ) and microscopic velocity  $\mathbf{v} \in \mathbb{R}^D$ . The Boltzmann collision operation  $\mathcal{Q}(F, F)$  models the binary interaction between the particles and takes the following form

$$\mathcal{Q}(F, F)(\mathbf{v}) = \int_{\mathbb{R}^D} \int_{\mathbb{S}^{D-1}} \mathcal{B}(\mathbf{v} - \mathbf{v}_*, \omega) [F(\mathbf{v}')F(\mathbf{v}'_*) - F(\mathbf{v})F(\mathbf{v}_*)] d\omega d\mathbf{v}_* \quad (1.2)$$

for the monatomic gases. Here

$$\mathbf{v}' = \frac{\mathbf{v} + \mathbf{v}_*}{2} + \frac{|\mathbf{v} - \mathbf{v}_*|}{2} \omega, \quad \mathbf{v}'_* = \frac{\mathbf{v} + \mathbf{v}_*}{2} - \frac{|\mathbf{v} - \mathbf{v}_*|}{2} \omega$$

---

\*Department of Mathematics, National University of Singapore, Singapore 119076, email: [matcz@nus.edu.sg](mailto:matcz@nus.edu.sg)

†Department of Mathematics, Stanford University, Stanford, CA 94305, email: [ywfan@stanford.edu](mailto:ywfan@stanford.edu)

‡Department of Mathematics and Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, email: [lexing@stanford.edu](mailto:lexing@stanford.edu)

are the post-collisional velocities of two particles with pre-collisional velocities  $\mathbf{v}$  and  $\mathbf{v}_*$ , and  $\omega$  is the angular parameter of the collision. The collision kernel  $\mathcal{B}$  is a non-negative function, and usually has the form

$$\mathcal{B}(\mathbf{v} - \mathbf{v}_*, \omega) = b(|\mathbf{v} - \mathbf{v}_*|, \cos \theta), \quad \cos \theta = |(\mathbf{v} - \mathbf{v}_*) \cdot \omega| / |\mathbf{v} - \mathbf{v}_*|.$$

Given that  $F(t, \mathbf{x}, \mathbf{v})$  denotes the mass density of particles, Boltzmann equation (1.1) guarantees  $F(t, \mathbf{x}, \mathbf{v})$  is non-negative if the initial value  $F(t = 0, \mathbf{x}, \mathbf{v})$  is non-negative. Noticing the symmetry of the collision (1.2) and  $d\mathbf{v} d\mathbf{v}_* = d\mathbf{v}' d\mathbf{v}'_*$ , we have formally the following equality for any function  $\psi(\cdot)$ :

$$\begin{aligned} \int_{\mathbb{R}^D} \psi(\mathbf{v}) \mathcal{Q}(F, F)(\mathbf{v}) d\mathbf{v} = \\ \frac{1}{4} \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} \int_{\mathbb{S}^{D-1}} (\psi(\mathbf{v}) + \psi(\mathbf{v}_*) - \psi(\mathbf{v}') - \psi(\mathbf{v}'_*)) \mathcal{B}[F(\mathbf{v}')F(\mathbf{v}'_*) - F(\mathbf{v})F(\mathbf{v}_*)] d\mathbf{v}_* d\mathbf{v} d\omega. \end{aligned} \quad (1.3)$$

By setting  $\psi(\mathbf{v}) = 1, \mathbf{v}, |\mathbf{v}|^2$ , one can derive the conservation of the mass, momentum and energy

$$\int_{\mathbb{R}^D} \mathcal{Q}(F, F)(\mathbf{v}) d\mathbf{v} = \int_{\mathbb{R}^D} \mathcal{Q}(F, F)(\mathbf{v}) \mathbf{v} d\mathbf{v} = \int_{\mathbb{R}^D} \mathcal{Q}(F, F)(\mathbf{v}) |\mathbf{v}|^2 d\mathbf{v} = 0. \quad (1.4)$$

The H-theorem

$$\int_{\mathbb{R}^D} \mathcal{Q}(F, F)(\mathbf{v}) \ln(F) d\mathbf{v} \leq 0, \quad (1.5)$$

obtained by setting  $\psi(\mathbf{v}) = \ln(F)$  states the monotonicity of the entropy.

Furthermore, by change of variables  $\mathbf{p} = \mathbf{v}' - \mathbf{v}$  and  $\mathbf{q} = \mathbf{v}'_* - \mathbf{v}$ , the Boltzmann collision operator can be rewritten as

$$\mathcal{Q}(F, F)(\mathbf{v}) = \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} \tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) [F(\mathbf{v} + \mathbf{p})F(\mathbf{v} + \mathbf{q}) - F(\mathbf{v})F(\mathbf{v} + \mathbf{p} + \mathbf{q})] d\mathbf{p} d\mathbf{q}. \quad (1.6)$$

This is the well-known as the Carleman form[5] of the Boltzmann collision operator where  $\tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q})$  is related to  $\mathcal{B}(\mathbf{v} - \mathbf{v}_*, \omega)$  by

$$\tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q}) = 2^{D-1} \mathcal{B} \left( \sqrt{|\mathbf{p}|^2 + |\mathbf{q}|^2}, \frac{|\mathbf{p}|}{\sqrt{|\mathbf{p}|^2 + |\mathbf{q}|^2}} \right) (|\mathbf{p}|^2 + |\mathbf{q}|^2)^{(2-D)/2}. \quad (1.7)$$

We refer the readers to the literature [5, 17, 14] for the details of the derivation.

Boltzmann equation is an accurate model in the gas dynamics, but its high dimensional nature and the complexity of the collision operator pose great difficulty for its numerical simulation. A classical method is the direct simulation of Monte Carlo [1], which uses simulation particles to mimic gas molecules and handles the collisions in a stochastic way. As a Monte Carlo method, it approximates the high-dimensional system efficiently, while the convergence order is low and the numerical solution often appears oscillatory.

With exponentially increasing of the computing power, in recent decades it becomes more and more feasible to solve the Boltzmann equation with deterministic methods. When considering the deterministic methods, the complex collision integral poses the most serious difficulty in the discretization. In this paper, we focus on the spatially homogeneous Boltzmann equation

$$\frac{\partial F}{\partial t} = \mathcal{Q}(F, F) \quad (1.8)$$

for convenience. In the past decades, several deterministic schemes for the Boltzmann equation have been developed. The two methods that attracted most attention are the discrete velocity method (DVM) [9, 21, 23, 17] and the Fourier spectral method (FSM) [2, 18, 19]. Next, we briefly review these two deterministic methods.

### 1.1 Discrete velocity method

In the discrete velocity method (DVM), it is assumed that the velocities of the particles belong to a finite set  $\Xi = \{\mathbf{v}_k\}_{k=1}^K \subset \mathbb{R}^D$ . In this paper, in order to simplify the discussion, we assume that the velocities  $\{\mathbf{v}_k\}$  belong to the lattice  $h\mathbb{Z}^D$ . More specifically, for each  $D$ -dimensional multi-index  $i = (i_1, \dots, i_D)$  with  $i_d = -n, \dots, n$ ,  $d = 1, \dots, D$ ,  $\mathbf{v}_i = (hi_1, \dots, hi_D)$ . Clearly  $K = N^D$  with  $N = 2n + 1$ . The distribution function  $F(t, \mathbf{v})$  is naturally discretized by its value  $f(t, \mathbf{v}_i)$  evaluated at the velocity point  $\mathbf{v}_i \in \Xi$ . The governing equations of  $f(t, \mathbf{v}_i)$  are

$$\frac{\partial f(t, \mathbf{v}_i)}{\partial t} = Q(\mathbf{v}_i) := \sum_{j,k,l} A_{ij}^{kl} (f(\mathbf{v}_k)f(\mathbf{v}_l) - f(\mathbf{v}_i)f(\mathbf{v}_j)), \quad (1.9)$$

where  $Q(\mathbf{v}_i)$  serves as an approximation of  $\mathcal{Q}(F, F)(\mathbf{v}_i)$ . To simplify the notations, we use  $\sum_i$  by  $\sum_{d=1}^D \sum_{i_d=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor}$  hereafter, if not specified otherwise. Here  $f(\mathbf{v}_k)$  is the abbreviation of  $f(t, \mathbf{v}_k)$ . The coefficients  $A_{ij}^{kl}$  are non-negative constants, and satisfy the conservation relation

$$A_{ij}^{kl} = \mathbf{1}(\mathbf{v}_i + \mathbf{v}_j - \mathbf{v}_k - \mathbf{v}_l) \mathbf{1}(|\mathbf{v}_i|^2 + |\mathbf{v}_j|^2 - |\mathbf{v}_k|^2 - |\mathbf{v}_l|^2) \Gamma_{ij}^{kl}, \quad (1.10)$$

and the symmetry relations

$$A_{ij}^{kl} = A_{ji}^{kl} = A_{kl}^{ij}. \quad (1.11)$$

Here  $\mathbf{1}(x)$  is the indicator function, i.e. it is equal to 1 if  $x = 0$  and 0 otherwise. The coefficients  $\Gamma_{ij}^{kl}$  depend on the collision kernel  $\mathcal{B}$  and the details of the discrete scheme used.

Due to its discrete nature, every collision involved in the DVM matches one of the admissible collisions in the continuous collision operator. This implies that the DVM maintains a number of fundamental physical properties of the continuous Boltzmann equation, such as positivity of the distribution function, the H-theorem and the exact conservation of mass, momentum and energy. Precisely speaking, the values  $f(t, \mathbf{v}_i)$  are always non-negative if the initial values  $f(t=0, \mathbf{v}_i)$  are non-negative. The symmetry relations of  $A_{ij}^{kl}$  implies that

$$\sum_i Q(\mathbf{v}_i) \psi(\mathbf{v}_i) = \frac{1}{4} \sum_{i,j,k,l} A_{ij}^{kl} (\psi(\mathbf{v}_i) + \psi(\mathbf{v}_j) - \psi(\mathbf{v}_k) - \psi(\mathbf{v}_l)) (f(\mathbf{v}_k)f(\mathbf{v}_l) - f(\mathbf{v}_i)f(\mathbf{v}_j)). \quad (1.12)$$

The conservation relations of  $A_{ij}^{kl}$  (1.10) indicate the conservation of mass, momentum and energy are preserved on a discrete level

$$\sum_i Q(\mathbf{v}_i) = \sum_i Q(\mathbf{v}_i) \mathbf{v}_i = \sum_i Q(\mathbf{v}_i) |\mathbf{v}_i|^2 = 0. \quad (1.13)$$

Recalling that the coefficients  $A_{ij}^{kl}$  are non-negative and letting  $\psi(\mathbf{v}_i) = \ln(f(\mathbf{v}_i))$ , one directly obtains a discrete version of the H-theorem of the discrete velocity method (DVM)

$$\sum_i Q(\mathbf{v}_i) \ln(f(\mathbf{v}_i)) \leq 0. \quad (1.14)$$

We emphasize here that the symmetry relations (1.11), non-negativity of  $A_{ij}^{kl}$  and the non-negativity of the initial values are key to the H-theorem.

On the other hand, as a result of using direct discretization for the high dimensional integration in the collision term, the DVM has a rather high computational cost  $O(N^{2D+\delta})$  for some  $0 < \delta \leq 1$ . It is also difficult to achieve high accuracy due to the insufficient collision pairs of velocities found in a Cartesian grid. More precisely, for the 2D case, the rate of convergence of the DVM in [9] is only  $O((1/\log h)^p)$  with  $p < 1/2 - 1/\pi$  [6]. And for 3D case, the best rate of convergence of the DVM is also lower than first order [16, 17].

## 1.2 Fourier spectral method

The Fourier spectral method assumes that the distribution function  $F$  has a compact support in  $\mathbf{v}$ :  $\text{Supp}(F) \subset B_S$ , where  $B_S$  is a ball centered at the origin with radius  $S$ . Then  $\text{Supp}(\mathcal{Q}(F, F)) \subset B_{\sqrt{2}S}$  and the collision term  $\mathcal{Q}(F, F)$  reduces to its truncated version  $Q(F, F)$  defined as

$$Q(F, F)(\mathbf{v}) := \int_{B_R} \int_{B_R} \tilde{B}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) [F(\mathbf{v} + \mathbf{p}) F(\mathbf{v} + \mathbf{q}) - F(\mathbf{v}) F(\mathbf{v} + \mathbf{p} + \mathbf{q})] d\mathbf{p} d\mathbf{q}. \quad (1.15)$$

Here we use the uppercase  $Q$  instead of the script font  $\mathcal{Q}$  to denote the truncated collision operator. In order to obtain a spectral approximation to the collision term, we restrict the distribution function  $F$  to the cube  $\mathcal{D}_T = [-T, T]^D$  with  $T \geq \frac{3\sqrt{2}+1}{4}R$  in order to avoid the aliasing and extend it periodically to the whole space (see [19, 14] for details). Then  $F(\mathbf{v})$  can be approximated by a truncated Fourier series

$$f(\mathbf{v}) := \sum_k \hat{f}_k E_k(\mathbf{v}), \quad \hat{f}_k = \frac{1}{(2T)^D} \int_{\mathcal{D}_T} F(\mathbf{v}) E_{-k}(\mathbf{v}) d\mathbf{v}, \quad (1.16)$$

where  $E_k(\mathbf{v}) = \exp(\frac{i\pi}{T} \mathbf{k} \cdot \mathbf{v})$ . Again here  $\sum_k$  stands for  $\sum_{d=1}^D \sum_{k_d=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor}$  hereafter, if not specified otherwise. By substituting (1.16) into (1.15) and performing a Galerkin projection, one can obtain for  $\hat{Q}_k$ , the discrete Fourier modes of  $Q(f, f)(\mathbf{v})$  on  $[-T, T]^D$ , the relationship

$$\hat{Q}_k = \sum_{l, m} \mathbf{1}(l + m - k) \left( \hat{B}(l, m) - \hat{B}(m, m) \right) \hat{f}_l \hat{f}_m, \quad (1.17)$$

where

$$\hat{B}(l, m) = \int_{B_R} \int_{B_R} \tilde{B}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) E_l(\mathbf{p}) E_m(\mathbf{q}) d\mathbf{p} d\mathbf{q}, \quad (1.18)$$

and  $\mathbf{1}(l)$  is indicator function as that in (1.10). It is easy to check that the coefficients  $\hat{B}(l, m)$  are real and satisfy the symmetry relations

$$\hat{B}(l, m) = \hat{B}(m, l) = \hat{B}(l, -m). \quad (1.19)$$

**Remark 1.** The above introduction to the Fourier spectral method is based on the Carleman representation of the Boltzmann collision operator (1.6). Starting from the classical form (1.2), we can also derive a relationship of form (1.17) (see [19] for details), while

the definition of  $\hat{B}(l, m)$  is slightly different:

$$\begin{aligned}\hat{B}(l, m) &= \int_{B_R} \int_{\mathbb{S}^{D-1}} \mathcal{B}(\mathbf{g}, \omega) E_l \left( \frac{1}{2}(\mathbf{g} + |\mathbf{g}|\omega) \right) E_m \left( \frac{1}{2}(\mathbf{g} - |\mathbf{g}|\omega) \right) d\omega d\mathbf{g} \\ &= \int_{B_R} \int_{B_R} \mathcal{B}(\mathbf{g}, \mathbf{g}'/|\mathbf{g}'|) \delta(|\mathbf{g}|^D - |\mathbf{g}'|^D) E_l \left( \frac{1}{2}(\mathbf{g} + \mathbf{g}') \right) E_m \left( \frac{1}{2}(\mathbf{g} - \mathbf{g}') \right) d\mathbf{g}' d\mathbf{g},\end{aligned}\tag{1.20}$$

where  $T \geq \frac{3+\sqrt{2}}{4}R$  and  $\mathbf{g}' = |\mathbf{g}|\omega$ . It is straightforward to check that these coefficients also have the symmetry relation (1.19).

Putting together the discussion so far, we arrive at the the following evolution equations of the discrete Fourier coefficients for the FSM [19]

$$\begin{cases} \frac{d\hat{f}_k}{dt} = \sum_{l,m} \mathbf{1}(l+m-k) \left( \hat{B}(l, m) - \hat{B}(m, m) \right) \hat{f}_l \hat{f}_m, \\ \hat{f}_k(t=0) = \hat{f}_k^0, \end{cases}\tag{1.21}$$

where  $\hat{f}_k^0$  are the truncated Fourier coefficients of  $F(t=0, \mathbf{v})$  restricted on  $[-T, T]^D$ . Due to the Fourier expansion, the FSM achieves the spectral accuracy, although the computational cost is still as high as  $O(N^{2D})$  [19]. Two different fast algorithms [14, 7] were proposed to accelerate the FSM, which reduced the computational cost to  $O(MN^D \log(N))$  for the hard sphere molecules (Maxwell molecules for 2D case) [14], and to  $O(MN^{D+1} \log(N))$  for general collision kernels [7], where  $M$  is the number of points to discretize the sphere  $\mathbb{S}^{D-1}$ . However, the solution of the FSM loses most of the aforementioned physical properties, including positivity, the H-theorem and the conservation of momentum and energy. In [20], Pareschi and Russo proposed a positivity preserving regularization of the FSM by using Fejér filter with the expense of losing spectral accuracy. However, their method fails to satisfy the H-theorem. The loss of the conservation was fixed in [8] by a spectral-Lagrangian modification. Unfortunately, these two regularizations cannot be hybridized to preserve both the positivity and the conservation at the same time since the spectral-Lagrangian method breaks the positivity.

### 1.3 Motivation

The classical DVM preserves a number of physical properties including positivity, the H-theorem and exact conservation of mass, momentum and energy, but suffers from high computational cost and low accuracy. The classical FSM has spectral accuracy and lower computational cost, but in order to gain the spectral accuracy, several key properties such as positivity, the H-theorem, and the conservation of momentum and energy are dropped. In this paper, we aim for a trade-off between the physical properties and the spectral accuracy.

The positivity of the solution a fundamental property of the solution, it helps to establish the H-theorem, which is one of the primary properties to guarantee the well-posedness of the discrete system. Therefore, it is worth sacrificing a bit the accuracy in exchange for positivity and the H-theorem. The goal of this paper is to obtain a spectral method preserving the positivity, mass conservation and the H-theorem, and with the same computational cost as the FSM. As for the accuracy, we are able to achieve the second order, which is acceptable in quite a number of applications.

To achieve this goal, we carefully study the reason behind the lack of the H-theorem for the classical FSM by a comparison with DVM. Then with some modification to the classical FSM, we propose an entropy monotonic spectral method (EMSM) that fulfills all the properties we expected.

The rest of the paper is organized as follows. In Section 2, we first outline the key steps to achieve this goal and state the main results of the paper. The details of the derivation of the EMSM is provided in Section 3. Section 4 presents the implementation of the EMSM and the numerical results. The paper ends with a discussion in Section 5.

## 2 Main result

In this section, we list our technical route and main results. Detailed discussion and investigation is given in Section 3.

To guarantee the H-theorem of the DVM, we need the following condition:

**Condition 1.** *The DVM satisfies*

1. *the coefficients  $A_{ij}^{kl}$  in (1.9) satisfy the symmetry relation  $A_{ij}^{kl} = A_{ji}^{kl} = A_{kl}^{ij}$ ;*
2. *the coefficients  $A_{ij}^{kl}$  in (1.9) are non-negative, i.e  $A_{ij}^{kl} \geq 0$ ;*
3. *the initial values is non-negative, i.e  $f(t=0, \mathbf{v}_i) \geq 0$ .*

Using the Fourier transforms

$$f(\mathbf{v}_p) = \sum_k \hat{f}_k E_k(\mathbf{v}_p), \quad \hat{f}_k = \frac{1}{ND} \sum_p f(\mathbf{v}_p) E_{-k}(\mathbf{v}_p), \quad (2.1)$$

where  $\mathbf{v}_p$  are corresponding collocation points in velocity space and  $k$  are the Fourier samples, one can write a scheme either in the velocity space in the DVM form or in the Fourier space in the FSM form.

In order to construct a scheme that satisfies the H-theorem, we follow the following three steps.

1. First, we modify the FSM (1.17) and rewrite the resulting scheme into the DVM form (1.9)

$$\tilde{Q}(\mathbf{v}_r) = \sum_{p,q,s} \tilde{A}_{rs}^{pq} [f(\mathbf{v}_p) f(\mathbf{v}_q) - f(\mathbf{v}_r) f(\mathbf{v}_s)], \quad (2.2)$$

where  $\tilde{A}_{rs}^{pq}$  are defined in (3.16) and satisfy the symmetry relation Condition 1.1 (see Section 3.1 for details). In this process, the collision term of the resulting FSM (1.17) also takes the following form in the Fourier domain

$$\hat{\tilde{Q}}_k = \sum_{l,m} \Delta(l+m-k) [\hat{B}(l, m) - \hat{B}(m, m)] \hat{f}_l \hat{f}_m, \quad (2.3)$$

where  $\Delta(l) := \mathbf{1}(l \bmod N)$ .

2. A careful study shows that  $\tilde{A}_{rs}^{pq}$  is not non-negative, but this can be fixed by applying a positivity preserving filter to  $\hat{B}(l, m)$ , i.e.,

$$\hat{B}^\sigma(l, m) := \hat{B}(l, m) \sigma_N(l) \sigma_N(m), \quad \sigma_N(l) = \prod_{d=1}^D \bar{\sigma}_N(l_d), \quad (2.4)$$

where  $\bar{\sigma}_N(n)$  is the modified Jackson filter [12, 24] given by

$$\bar{\sigma}_N(n) = \frac{(m+1-|n|) \cos(\frac{\pi|k|}{m+1}) + \sin(\frac{\pi|k|}{m+1}) \cot(\frac{\pi}{m+1})}{m+1}, \quad m = \lfloor \frac{N-1}{2} \rfloor. \quad (2.5)$$

Accordingly, the coefficients  $\tilde{A}_{rs}^{pq}$  are modified to become  $\tilde{\tilde{A}}_{rs}^{pq}$  (see (3.41)), which satisfy the symmetry relation (Condition 1.1) and are non-negative (Condition 1.2).

3. To guarantee the positivity of the initial values (Condition 1.3), we adopt interpolation rather than orthogonal projection while generating the initial condition, i.e.  $f^I(t=0, \mathbf{v}_p) = F(t=0, \mathbf{v}_p)$  for any  $\mathbf{v}_p$  if  $F(t=0, \mathbf{v})$  is smooth, otherwise see (3.44) for details.

**Main result.** Summarizing the outline given above, we arrive at a new entropy monotonic spectral method (EMSM) that takes the following simple form

$$\begin{cases} \frac{d\hat{f}_k}{dt} = \sum_{l,m} \Delta(l+m-k) \left( \hat{B}^\sigma(l, m) - \hat{B}^\sigma(m, m) \right) \hat{f}_l \hat{f}_m, \\ \hat{f}_k(t=0) = \hat{f}_k^{I,0}, \end{cases} \quad (2.6)$$

where  $\hat{f}_k^{I,0}$  are the discrete Fourier coefficients of  $f^I(t=0, \mathbf{v})$ . This method preserves several key physical properties, including positivity, the H-theorem, and mass conservation. Due to the positivity-preserving filter (2.5), the accuracy of EMSM is second order.

The fast algorithms proposed in [14, 7] can also be applied without much change to the EMSM. More precisely, in [14, 7]  $\hat{B}(l, m)$  is approximated by

$$\hat{B}(l, m) \approx \sum_{p=1}^P \alpha_{l+m}^{(p)} \beta_l^{(p)} \gamma_m^{(p)}. \quad (2.7)$$

Then  $\hat{B}^\sigma(l, m)$  can be approximated by

$$\hat{B}^\sigma(l, m) \approx \sum_{p=1}^P \alpha_{l+m}^{(p)} \left( \sigma_N(l) \beta_l^{(p)} \right) \left( \sigma_N(m) \gamma_m^{(p)} \right). \quad (2.8)$$

We claim that the fast algorithms do not destroys any physical properties of the EMSM, and the rigorous discussion is given in Section 4.1. As a result, the computational cost of the EMSM is same as that of FSM.

### 3 Entropy monotonic spectral method

As shown in Section 1.1, a discrete H-theorem can be obtained from the classical DVM, where the associated entropy function can be considered as a numerical quadrature for the integral of  $f \ln f$ . Obviously, this requires the positivity of the distribution function, which can be guaranteed by the positivity of the discrete collision kernel  $A_{ij}^{kl}$ . In general, to preserve the Boltzmann entropy in the numerical scheme, the positivity of the numerical solution needs to be enforced in a certain sense due to the presence of  $\ln f$  in the entropy

---

<sup>1</sup>If  $N$  is even, set  $\bar{\sigma}_N(-N/2) = 0$ .

function. However, in the FSM, there is no guarantee of any form for positivity in the numerical solution, and hence the H-theorem does not hold. In [20], the authors proposed a positivity preserving Fourier spectral method by introducing strong filters, so that the distribution function is point-wise non-negative and the approximation of the Boltzmann entropy becomes possible. However, while the numerical solution of this approach turns out to be highly dissipative [20], an H-theorem still seems to be missing. In this paper, rather than enforcing the non-negativity of the whole distribution function, we focus on its non-negativity on the collocation points. Due to the one-to-one mapping between the Fourier coefficients and the values at the collocation points, the Fourier spectral method can be rewritten as a scheme of function values defined at the collocation points, which appears to be a type of DVM. As stated at the end of Section 2, the H-theorem of DVM can be derived from the three conditions in Condition 1. Below we modify the FSM so that these three conditions are fulfilled.

This section is an detailed discussion of the Step 1-3 in Section 2. Section 3.1 corresponds to the Step 1, while Step 2 and 3 are studied in Section 3.2. Section 3.3 reviews and discusses all the methods referred in this section.

### 3.1 Symmetric Fourier spectral methods

As the first step, we start by revising the original FSM so that it satisfies the symmetry condition. Below we introduce two symmetric methods. The first version aims to match the gain term and the loss term in the DVM form. This version is further modified to give a second symmetric version so that the positivity can be more easily realized later on.

#### 3.1.1 A symmetric FSM by altering the loss term

Specializing the definition of  $f(\mathbf{v})$  in the FSM (1.16) to the sampled velocities  $\mathbf{v}_p$ , we arrive at

$$f(\mathbf{v}_p) = \sum_k \hat{f}_k E_k(\mathbf{v}_p), \quad \hat{f}_k = \frac{1}{N^D} \sum_p f(\mathbf{v}_p) E_{-k}(\mathbf{v}_p). \quad (3.1)$$

The collision term  $Q(\mathbf{v}) := \sum_k \hat{Q}_k E_k(\mathbf{v})$  can be calculated from (1.17). By denoting  $Q^+$  and  $Q^-$  the gain term and the loss term, respectively, we see that the values of the gain term on the points  $\mathbf{v}_r$  are

$$\begin{aligned} Q^+(\mathbf{v}_r) &= \sum_{l,m,k} \mathbf{1}(l+m-k) \hat{B}(l,m) \hat{f}_l \hat{f}_m E_k(\mathbf{v}_r) \\ &= \frac{1}{N^{2D}} \sum_{\substack{l,m,k \\ p,q}} \mathbf{1}(l+m-k) \hat{B}(l,m) f(\mathbf{v}_p) f(\mathbf{v}_q) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r). \end{aligned} \quad (3.2)$$

Notice that

$$\frac{1}{N^D} \sum_s E_j(\mathbf{v}_s) = \Delta(j) \quad (3.3)$$

where  $\Delta(j) := \mathbf{1}(j \bmod N)$ . Summing over  $j$  gives  $\frac{1}{N^D} \sum_{j,s} E_j(\mathbf{v}_s) = 1$ . Plugging this into (3.2) and applying (3.3) results in

$$\begin{aligned} Q^+(\mathbf{v}_r) &= \frac{1}{N^{3D}} \sum_{\substack{l,m,k,j \\ p,q,s}} \mathbf{1}(l+m-k) \mathbf{1}(j) \hat{B}(l,m) f(\mathbf{v}_p) f(\mathbf{v}_q) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s) \\ &= \frac{1}{N^{3D}} \sum_{\substack{l,m,k,j \\ p,q,s}} \mathbf{1}(l+m-k-j) \hat{B}(l-j, m-j) f(\mathbf{v}_p) f(\mathbf{v}_q) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s). \end{aligned} \quad (3.4)$$



Here we extend the definition of  $\hat{B}$  periodically modulus  $N$  in each dimension, i.e.,  $\hat{B}(l - j, m - j) = \hat{B}((l - j) \bmod N, (m - j) \bmod N)$ . If we introduce

$$A_{rs}^{pq} = \frac{1}{N^{3D}} \sum_{l,m,k,j} \mathbf{1}(l + m - k - j) \hat{B}(l - j, m - j) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s), \quad (3.5)$$

then the gain term is written as

$$Q^+(\mathbf{v}_r) = \sum_{p,q,s} A_{rs}^{pq} f(\mathbf{v}_p) f(\mathbf{v}_q). \quad (3.6)$$

Apparently, such term does take the form of a gain term in the DVM (1.9). Thus, in order to enforce symmetry of the discrete collision operator in the DVM form, the loss term needs to be changed to

$$\tilde{Q}^-(\mathbf{v}_r) = \sum_{p,q,s} A_{rs}^{pq} f(\mathbf{v}_r) f(\mathbf{v}_s), \quad (3.7)$$

and thus the whole collision term reads

$$\tilde{Q}(\mathbf{v}_r) = \sum_{p,q,s} A_{rs}^{pq} [f(\mathbf{v}_p) f(\mathbf{v}_q) - f(\mathbf{v}_r) f(\mathbf{v}_s)]. \quad (3.8)$$

Recall that the first part of Condition 1 is the symmetry of the coefficients. We need to show that

$$A_{rs}^{pq} = A_{sr}^{pq} = A_{pq}^{rs}. \quad (3.9)$$

These equalities follow from the result of the symmetry of  $\hat{B}(l, m)$  in (1.19).

In order to justify the modified loss term  $\tilde{Q}^-(\mathbf{v}_r)$ , we notice that from (3.3) one has

$$\sum_{p,q} A_{rs}^{pq} = \frac{1}{N^D} \sum_{k,j} \mathbf{1}(k + j) \hat{B}(j, j) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s). \quad (3.10)$$

The modified loss term is

$$\begin{aligned} \tilde{Q}^-(\mathbf{v}_r) &= \frac{1}{N^D} \sum_{\substack{k,j \\ s}} \mathbf{1}(k + j) \hat{B}(j, j) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s) f(\mathbf{v}_r) f(\mathbf{v}_s) \\ &= \sum_{k,j} \mathbf{1}(k + j) \hat{B}(j, j) E_k(\mathbf{v}_r) f(\mathbf{v}_r) \hat{f}_{-j} \\ &= f(\mathbf{v}_r) \sum_k \hat{B}(k, k) E_k(\mathbf{v}_r) \hat{f}_k. \end{aligned} \quad (3.11)$$

Actually, the original loss term  $Q^-$  in (1.15) can be written as

$$Q^-(\mathbf{v}) = f(\mathbf{v}) \mathcal{L}(f)(\mathbf{v}), \quad \mathcal{L}(f)(\mathbf{v}) = \int_{B_R} \int_{B_R} \tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q}) f(\mathbf{v} + \mathbf{p} + \mathbf{q}) d\mathbf{p} d\mathbf{q}. \quad (3.12)$$

One can obtain (3.11) by applying the Fourier-collocation spectral method to (3.12). In this sense, the symmetric collision term (3.8) is a hybrid Fourier approximation of the Boltzmann equation, with the Fourier-Galerkin spectral method used for the gain term and the Fourier-collocation spectral method for the loss term.

Moreover, using (3.1), we can calculate the Fourier coefficients of  $\tilde{Q}^-$ ,

$$\begin{aligned}\hat{\tilde{Q}}_k^- &= \frac{1}{N^D} \sum_r \tilde{Q}^-(\mathbf{v}_r) E_{-k}(\mathbf{v}_r) \\ &= \frac{1}{N^D} \sum_{r,l,m} \hat{B}(m,m) E_m(\mathbf{v}_r) \hat{f}_m \hat{f}_l E_l(\mathbf{v}_r) E_{-k}(\mathbf{v}_r) \\ &= \sum_{l,m} \Delta(l+m-k) \hat{B}(m,m) \hat{f}_l \hat{f}_m,\end{aligned}\tag{3.13}$$

where (3.3) is used in the last equality. Notice that (3.13) can be calculated with an FFT based convolution. The modification (3.11) essentially changes the anti-aliasing convolution (loss term of (1.17)) into a classical periodic convolution (3.13). With the new loss term, the Fourier coefficient of  $E_k(\mathbf{v})$  in the new discrete collision term reads

$$\sum_{l,m} [\mathbf{1}(l+m-k) \hat{B}(l,m) - \Delta(l+m-k) \hat{B}(m,m)] \hat{f}_l \hat{f}_m.\tag{3.14}$$

### 3.1.2 A symmetric FSM with better symmetry

Comparing (3.14) and (1.17), we find that in the Fourier space form, the new collision operator is even less natural due to the appearance of both  $\mathbf{1}$  and  $\Delta$ . Therefore it is natural to replace the gain term as its “aliased” counterpart:

$$\hat{\tilde{Q}}_k^+ = \sum_{l,m} \Delta(l+m-k) \hat{B}(l,m) \hat{f}_l \hat{f}_m.\tag{3.15}$$

It is interesting that a combination of (3.15) and (3.13) still gives a symmetric collision term if  $N$  is odd, and therefore the resulting collision term in the velocity space has symmetry in terms of both the Fourier coefficients and the values at the collocation points. For the gain term, repeating the procedure in (3.2)-(3.6) and defining

$$\tilde{A}_{rs}^{pq} = \frac{1}{N^{3D}} \sum_{j,k,l,m} \Delta(l+m-j-k) \hat{B}(l-j, m-j) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s),\tag{3.16}$$

results in

$$\tilde{Q}^+(\mathbf{v}_r) = \sum_{p,q,s} \tilde{A}_{rs}^{pq} f(\mathbf{v}_p) f(\mathbf{v}_q).\tag{3.17}$$

Here we use the fact that both  $\Delta$  and  $\hat{B}$  are periodic modulus  $N$ .

For the loss term, we notice that

$$\begin{aligned}\sum_{p,q} \tilde{A}_{rs}^{pq} &= \frac{1}{N^D} \sum_{j,k} \Delta(-j-k) \hat{B}(-j, -j) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s) \\ &= \frac{1}{N^D} \sum_{j,k} \mathbf{1}(j+k) \hat{B}(j, j) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s) = \sum_{p,q} A_{rs}^{pq}.\end{aligned}\tag{3.18}$$

Here we have  $\Delta(-j-k) = \mathbf{1}(j+k)$  since every component of  $j$  and  $k$  ranges from  $-(N-1)/2$  to  $(N-1)/2$ . Therefore,

$$\tilde{Q}^-(\mathbf{v}_r) = \sum_{p,q,s} \tilde{A}_{rs}^{pq} f(\mathbf{v}_r) f(\mathbf{v}_s).\tag{3.19}$$

Thus, the whole discrete collision term

$$\hat{Q}_k = \sum_{l,m} \Delta(l+m-k) [\hat{B}(l,m) - \hat{B}(m,m)] \hat{f}_l \hat{f}_m \quad (3.20)$$

is also symmetric in the DVM form. Similar to the symmetry relation (3.9) of  $A_{rs}^{pq}$ , one can obtain

$$\tilde{A}_{rs}^{pq} = \tilde{A}_{sr}^{pq} = \tilde{A}_{pq}^{rs}. \quad (3.21)$$

We point out that the modification to the collision term is equivalent to approximating  $Q(f, f)$  by collocation method rather Galerkin method. Hence, this modification does not ruin the spectral accuracy.

If  $N$  is even, the equality (3.18) does not hold since  $\Delta(-j-k) = 1$  but  $\mathbf{1}(j+k) = 0$  if there exists a  $d = 1, \dots, D$  such that  $j_d = k_d = -N/2$  and  $j_{d'} + k_{d'} \bmod N = 0$  for all  $d' \neq d$ . However, the symmetry can be recovered by applying a filter  $\sigma_N(l, m)$  to  $\hat{B}(l, m)$ , so that the kernel  $\hat{B}(l, m)$  is replaced by  $\sigma_N(l, m)\hat{B}(l, m)$ . If  $\sigma_N(l, m)$  satisfies

$$\sigma_N(l, m) = \sigma_N(m, l) = \sigma_N(-l, m), \quad \sigma_N(j, k) = 0 \text{ if } \min(j_1, \dots, j_D, k_1, \dots, k_D) = -N/2, \quad (3.22)$$

then (3.18) holds again, and the symmetry follows. In this case, the accuracy of the spectral method depends on the filter.

**Remark 2.** In the literature, the symmetric FSM (3.21) has already appeared in some fast summation algorithms (c.f. [14, 7]). For example, in [14], the following decomposition of  $\hat{B}(l, m)$  is considered:

$$\hat{B}(l, m) \approx \sum_{p=1}^P \beta_l^{(p)} \gamma_m^{(p)}, \quad (3.23)$$

where  $P \in \mathbb{N}^+$  is the total number of quadrature points on the sphere. Then the collision term can be approximated by

$$\hat{Q}_k \approx \sum_{p=1}^P \sum_{l,m} \mathbf{1}(l+m-k) \left[ \left( \beta_l^{(p)} \hat{f}_l \right) \left( \gamma_m^{(p)} \hat{f}_m \right) - \hat{f}_l \left( \beta_m^{(p)} \gamma_m^{(p)} \hat{f}_m \right) \right]. \quad (3.24)$$

To evaluate (3.24) efficiently, one needs to utilize FFT-based convolutions. In the implementation, if no anti-aliasing technique is used, the summation over  $l, m$  would actually be subject to  $k = (l+m) \bmod N$ , which is identical to (3.21). In practical implementation, when  $N$  is large, an implementation without anti-aliasing is also acceptable, and such techniques are not mentioned in [14, 7].

## 3.2 A positive and entropic Fourier spectral method

As we remarked earlier, in order to obtain an H-theorem for a symmetric DVM, one needs to ensure that all the coefficients  $\tilde{A}_{pq}^{rs}$  are non-negative (Condition 1.2). Below, we first show that  $\tilde{A}_{pq}^{rs}$  fail to be non-negative, and then apply a filter to recover non-negativity.

### 3.2.1 Failure of positivity preservation in the symmetric FSM

A straightforward calculation yields

$$\begin{aligned} \tilde{A}_{rs}^{pq} &= \frac{1}{N^{3D}} \sum_{l,m,k,j} \Delta(l+m-k-j) \hat{B}(l-j, m-j) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r) E_j(\mathbf{v}_s) \\ &= \frac{1}{N^{3D}} \sum_{l,m,k} \hat{B}(m-k, l-k) E_{-l}(\mathbf{v}_p - \mathbf{v}_s) E_{-m}(\mathbf{v}_q - \mathbf{v}_s) E_k(\mathbf{v}_r - \mathbf{v}_s). \end{aligned} \quad (3.25)$$

Thus, we have  $\tilde{A}_{rs}^{pq} = \tilde{A}_{r-s,0}^{p-s,q-s}$ . Since

$$\begin{aligned}\tilde{A}_{r0}^{pq} &= \frac{1}{N^{3D}} \sum_{l,m,k} \hat{B}(m-k, l-k) E_{-l}(\mathbf{v}_p) E_{-m}(\mathbf{v}_q) E_k(\mathbf{v}_r) \\ &= \frac{1}{N^{3D}} \sum_{i,j,k} \hat{B}(i, j) E_{-k-i}(\mathbf{v}_p) E_{-k-j}(\mathbf{v}_q) E_k(\mathbf{v}_r) \\ &= \Delta(r-p-q) \frac{1}{N^{2D}} \sum_{i,j} \hat{B}(i, j) E_{-i}(\mathbf{v}_p) E_{-j}(\mathbf{v}_q).\end{aligned}\tag{3.26}$$

we arrive at

$$\tilde{A}_{rs}^{pq} = \Delta(r+s-p-q) \frac{1}{N^{2D}} \sum_{l,m} \hat{B}(l, m) E_{-l}(\mathbf{v}_p - \mathbf{v}_s) E_{-m}(\mathbf{v}_q - \mathbf{v}_s).\tag{3.27}$$

Let

$$G(\mathbf{p}', \mathbf{q}') = \frac{1}{N^{2D}} \sum_{l,m} \hat{B}(l, m) E_{-l}(\mathbf{p}') E_{-m}(\mathbf{q}').\tag{3.28}$$

A direct calculation yields

$$G(\mathbf{p}', \mathbf{q}') = \int_{B_R} \int_{B_R} \tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) \chi_N(\mathbf{p} - \mathbf{p}') \chi_N(\mathbf{q} - \mathbf{q}') d\mathbf{p} d\mathbf{q},\tag{3.29}$$

where

$$\chi_N(\mathbf{v}) = \frac{1}{N^D} \sum_k E_k(\mathbf{v}).\tag{3.30}$$

Let

$$K(\mathbf{p}, \mathbf{q}) = \tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) \mathbf{1}(|\mathbf{p}| \leq R) \mathbf{1}(|\mathbf{q}| \leq R), \quad \mathbf{p}, \mathbf{q} \in [-T, T]^D,\tag{3.31}$$

and extend it to the whole space periodically with period  $2T$ . Apparently,  $G(\mathbf{p}', \mathbf{q}')$  is the truncated Fourier expansion of  $K(\mathbf{p}, \mathbf{q})$ .

Using the definition of  $G$ , one notices that  $\tilde{A}_{rs}^{pq} = \Delta(r+s-p-q)G(\mathbf{v}_p - \mathbf{v}_s, \mathbf{v}_q - \mathbf{v}_s)$ . Therefore, the coefficients  $\tilde{A}_{rs}^{pq}$  are all non-negative if and only if  $G(\mathbf{v}_p, \mathbf{v}_q) \geq 0$  for all multi-indices  $p$  and  $q$ . However, this is generally violated due to the oscillatory behavior of  $\chi_N(\cdot)$ . Some values of  $G$  for the kernel  $\tilde{\mathcal{B}}(\mathbf{p}, \mathbf{q}) \equiv \frac{1}{\pi}$ ,  $R = 6$  and  $D = 2$  are plotted in Figure 1. This clearly shows that negative values appear regularly. Therefore in general, the H-theorem does not hold for the symmetric FSM proposed in Section 3.1.2.

### 3.2.2 Realizing H-theorem by filtering

In (3.29), one can see that if  $\chi_N(\cdot)$  were a non-negative function, then  $G(\mathbf{p}', \mathbf{q}')$  would be non-negative for any  $\mathbf{p}'$  and  $\mathbf{q}'$ . Thus, in order to get non-negative coefficients, we replace the function  $\chi_N$  by a non-negative one. Note that for any  $\mathbf{v}$ , we have the following limit:

$$\lim_{N \rightarrow +\infty} \chi_N(\mathbf{v}) = \chi(\mathbf{v}) := \mathbf{1}(\mathbf{v} \bmod 2T).\tag{3.32}$$

Therefore, in order to keep the consistency of the scheme, our aim is to find  $\chi_N^\sigma(\mathbf{v}) \geq 0$  which also has the limit  $\chi(\mathbf{v})$  as  $N \rightarrow +\infty$ .

As  $N \rightarrow +\infty$ , the function  $\chi_N(\cdot)$  tends to  $\chi(\mathbf{v})$  in an oscillatory behavior. Although the oscillation becomes weaker as  $N$  increases, negative values always appear when the

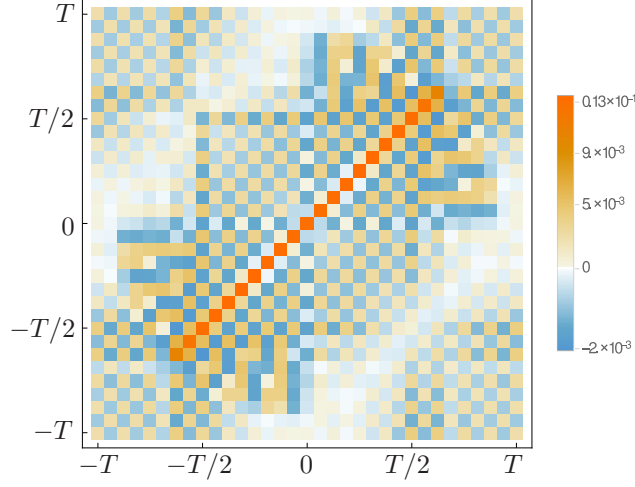


Figure 1: The values of  $G(\mathbf{v}_p, \mathbf{v}_q)$  with  $N = 32$  and  $\mathbf{v}_q = (T/2, T/2)$ . The axes are the two components of  $\mathbf{v}_p$ .

function oscillates around the value zero. In the spectral method, a common way to reduce the oscillation is filtering. Generally, for any positive integer  $N$  and  $T \in \mathbb{R}^+$ , we define the space

$$\mathbb{P}_N = \text{span} \left\{ E_k(\mathbf{v}) \mid -\left\lfloor \frac{N}{2} \right\rfloor \leq k_d \leq \left\lfloor \frac{N-1}{2} \right\rfloor, d = 1, \dots, D \right\} \subset L^2_{\text{per}}([-T, T]^D). \quad (3.33)$$

Thus a filter  $\mathcal{S}_N^\sigma$  is a linear map from  $L^2_{\text{per}}([-T, T]^D)$  to  $\mathbb{P}_N$ , which is defined as

$$\mathcal{S}_N^\sigma : \sum_{k \in \mathbb{Z}^D} \hat{f}_k E_k(\mathbf{v}) \mapsto \sum_k \sigma_N(k) \hat{f}_k E_k(\mathbf{v}). \quad (3.34)$$

Here we only consider the symmetric filters, i.e.,  $\sigma_N(k)$  must satisfy

$$\begin{aligned} \sigma_N(k) &= \prod_{d=1}^D \bar{\sigma}_N(k_d), \quad \bar{\sigma}_N(k_d) = \bar{\sigma}_N(-k_d), \quad \bar{\sigma}_N(0) = 1, \\ 0 \leq \bar{\sigma}_N(|k_d|) &\leq \bar{\sigma}_N(|k_d| - 1), \quad \lim_{N \rightarrow +\infty} \bar{\sigma}_N(k_d) = 1. \end{aligned} \quad (3.35)$$

Apparently  $\chi_N \in \mathbb{P}_N$ . If  $\chi_N^\sigma := \mathcal{S}_N^\sigma \chi_N$  is a positive function, then  $\chi_N^\sigma$  can be adopted as the replacement of  $\chi_N$  in the spectral method.

For any  $\sigma_N$  satisfying (3.35), if  $N$  is odd, we have that

$$\chi_N^\sigma(\mathbf{v}) = \frac{1}{N^D} \sum_k \sigma_N(k) E_k(\mathbf{v}) = \frac{1}{N^D} \prod_{d=1}^D \left( 1 + 2 \sum_{n=1}^{(N-1)/2} \bar{\sigma}_N(n) \cos(n\pi \xi_d/T) \right). \quad (3.36)$$

Now it is clear that any function  $\bar{\sigma}_N(\cdot)$  satisfying (3.35) and

$$1 + 2 \sum_{n=1}^{(N-1)/2} \bar{\sigma}_N(n) \cos(nx) \geq 0 \quad (3.37)$$

can be used to construct the filter. Especially, the modified Jackson filter [12, 24]

$$\bar{\sigma}_N(n) = \frac{(m+1-|n|)\cos(\frac{\pi|k|}{m+1}) + \sin(\frac{\pi|k|}{m+1})\cot(\frac{\pi}{m+1})}{m+1}, \quad m = \frac{N-1}{2} \quad (3.38)$$

with  $n = -m, \dots, m$  satisfies (3.37) (see [24] for example for a proof). If  $N$  is even, one can define

$$\bar{\sigma}_N(n) = \bar{\sigma}_{N-1}(n), \quad n = -N/2 + 1, \dots, N/2 - 1, \quad \bar{\sigma}_N(\pm N/2) = 0. \quad (3.39)$$

Since  $\bar{\sigma}_N(N/2) = 0$ , one also gets  $\chi_N^\sigma(\mathbf{v}) \geq 0$  with the same method.

Once  $\chi_N$  is replaced by  $\chi_N^\sigma$  in (3.29), the function  $G(\mathbf{p}', \mathbf{q}')$  is then revised to be

$$\begin{aligned} \tilde{G}(\mathbf{p}', \mathbf{q}') &= \int_{B_R} \int_{B_R} \tilde{B}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) \chi_N^\sigma(\mathbf{p} - \mathbf{p}') \chi_N^\sigma(\mathbf{q} - \mathbf{q}') d\mathbf{p} d\mathbf{q} \\ &= \frac{1}{N^{2D}} \sum_{l,m} \int_{B_R} \int_{B_R} \tilde{B}(\mathbf{p}, \mathbf{q}) \delta(\mathbf{p} \cdot \mathbf{q}) \sigma_N(l) \sigma_N(m) E_l(\mathbf{p} - \mathbf{p}') E_m(\mathbf{q} - \mathbf{q}') d\mathbf{p} d\mathbf{q} \\ &= \frac{1}{N^{2D}} \sum_{l,m} [\sigma_N(l) \sigma_N(m) \hat{B}(l, m)] E_{-l}(\mathbf{p}') E_{-m}(\mathbf{q}'), \end{aligned} \quad (3.40)$$

and we have  $\tilde{G}(\mathbf{p}', \mathbf{q}') \geq 0$ . Furthermore, we can mimic (3.27) and define

$$\tilde{A}_{rs}^{pq} = \Delta(r + s - p - q) \tilde{G}(\mathbf{v}_p - \mathbf{v}_s, \mathbf{v}_q - \mathbf{v}_s). \quad (3.41)$$

For an odd  $N$ , it is clear that these coefficients are symmetric and non-negative. For an even  $N$ , we recall the condition (3.22). The last line of (3.40) shows that applying the filter to  $\chi_N$  is equivalent to applying the filter to  $\hat{B}(l, m)$ . It is easy to verify that the modified Jackson filter (3.38)(3.39) satisfies (3.22), and thus the coefficients  $\tilde{A}_{rs}^{pq}$  are also symmetric and non-negative for even  $N$ . We emphasize here that the symmetry of  $\tilde{A}_{rs}^{pq}$  guarantees the conservation of mass (see Section 1.1 for details).

To get the H-theorem, we still need to ensure that the initial data are non-negative. Note that we only need the initial discrete distribution function to be non-negative at the collocation points, and therefore it is natural to use interpolation rather than orthogonal projection while preparing the initial data. Here the interpolation is defined by

$$\mathcal{I}_N : L^2([-T, T]^D) \rightarrow \mathbb{P}_N \quad (3.42)$$

such that if  $F \in H_2^m([-T, T]^D)$ ,  $m > D/2$ , then

$$(\mathcal{I}_N F)(\mathbf{v}_k) = F(\mathbf{v}_k), \quad \mathbf{v}_k = \frac{k}{N} 2T, \quad k_d = -\lfloor N/2 \rfloor, \dots, \lfloor (N-1)/2 \rfloor, \quad (3.43)$$

otherwise

$$(\mathcal{I}_N F)(\mathbf{v}_k) = g_\epsilon(\mathbf{v}_k), \quad g_\epsilon = \varphi_\epsilon * F, \quad (3.44)$$

where  $\varphi_\epsilon \geq 0$  is a mollifier, such that  $\|F - g_\epsilon\|_{L^2} < \epsilon$  small enough. Denoting  $f^I = \mathcal{I}_N F$ , the new collision term reads

$$Q^\sigma(f^I, f^I)(\mathbf{v}_r) = \sum_{p,q,s} \tilde{A}_{rs}^{pq} [f^I(\mathbf{v}_p) f^I(\mathbf{v}_q) - f^I(\mathbf{v}_r) f^I(\mathbf{v}_s)]. \quad (3.45)$$

By the same techniques used in (3.4), (3.25) and (3.26), a simplification yields

$$\hat{Q}_k^\sigma = \sum_{l,m} \Delta(l+m-k) \left( \hat{B}^\sigma(l,m) - \hat{B}^\sigma(m,m) \right) \hat{f}_l^I \hat{f}_m^I \quad (3.46)$$

in the Fourier space, where

$$\hat{B}^\sigma(l,m) = \hat{B}(l,m) \sigma_N(l) \sigma_N(m), \quad (3.47)$$

$$\hat{Q}_k^\sigma = \frac{1}{N^D} \sum_r Q^\sigma(\mathbf{v}_r) E_{-k}(\mathbf{v}_r), \quad \hat{f}_l^I = \frac{1}{N^D} \sum_p f^I(\mathbf{v}_p) E_{-l}(\mathbf{v}_p). \quad (3.48)$$

The fact that  $\hat{Q}_k^\sigma = 0$  ensure the mass conservation. Altogether, the Cauchy problem of the entropy monotonic spectral method (EMSM) reads

$$\begin{cases} \frac{df^I}{dt} = Q^\sigma(f^I, f^I), \\ f^I(t=0, \mathbf{v}) = (\mathcal{I}_N F)(t=0, \mathbf{v}). \end{cases} \quad (3.49)$$

**Remark 3.** The same technique can be applied to the Fourier spectral method derived from the classical form of the Boltzmann collision operator (1.2) to get an entropy preserving method. In the derivation of symmetric FSMs in Section 3.1, we only used the symmetry of  $\hat{B}(l,m)$ , which also holds for the coefficients defined in (1.20). Thus, by defining  $G(\mathbf{p}', \mathbf{q}')$  as (3.28), and noticing (1.20), one can directly obtain

$$G(\mathbf{p}', \mathbf{q}') = \int_{B_R} \int_{B_R} K(\mathbf{g}, \mathbf{g}') \chi_N \left( \frac{\mathbf{g} + \mathbf{g}'}{2} - \mathbf{p}' \right) \chi_N \left( \frac{\mathbf{g} - \mathbf{g}'}{2} - \mathbf{q}' \right) d\mathbf{g} d\mathbf{g}', \quad (3.50)$$

$$K(\mathbf{g}, \mathbf{g}') = \mathcal{B}(\mathbf{g}, \mathbf{g}' / |\mathbf{g}'|) \delta(|\mathbf{g}|^D - |\mathbf{g}'|^D) \mathbf{1}(|\mathbf{g}| \leq R) \mathbf{1}(|\mathbf{g}'| \leq R), \quad \mathbf{g}, \mathbf{g}' \in [-T, T]^D. \quad (3.51)$$

Again, by (3.26) (which also depends only on the symmetry of  $\hat{B}(l,m)$ ), the positivity of  $\hat{A}_{rs}^{pq}$  depends only on the positivity of  $G(\mathbf{p}', \mathbf{q}')$  at the collocation points. Therefore, replacing  $\chi_N$  with  $\chi_N^\sigma$  does the job.

### 3.3 Relation between different FSMs

As a summary, here we compare all different version of FSMs in this section. To set up a uniform notation, let  $\mathcal{P}_N$  be the projection operator from  $L_{\text{per}}^2([-T, T]^D)$  onto  $\mathbb{P}_N$  and  $\mathcal{G}(\cdot; \cdot, \cdot)$  be the general collision operator

$$\mathcal{G}(C; f, f)(\mathbf{v}) = \int_{[-T, T]^D} \int_{[-T, T]^D} C(\mathbf{p}, \mathbf{q}) [f(\mathbf{v} + \mathbf{p}) f(\mathbf{v} + \mathbf{q}) - f(\mathbf{v}) f(\mathbf{v} + \mathbf{p} + \mathbf{q})] d\mathbf{p} d\mathbf{q} \quad (3.52)$$

with a collision kernel  $C(\cdot, \cdot)$ . Thus the collision term (1.15) can be written as  $\mathcal{G}(K; f, f)$  using the definition of  $K$  in (3.31). Similar to  $Q^+$  and  $Q^-$ , the symbols  $\mathcal{G}^+$  and  $\mathcal{G}^-$  stand respectively for the corresponding gain term and loss term.

In our previous discussion, five different approximations of  $\mathcal{G}(K; f, f)$  and the initial values are involved:

$$\mathcal{P}_N \mathcal{G}^+(\mathcal{P}_N K; f, f) + \mathcal{P}_N \mathcal{G}^-(\mathcal{P}_N K; f, f), \quad f(t=0, \mathbf{v}) = \mathcal{P}_N F(t=0, \mathbf{v}), \quad (3.53)$$

$$\mathcal{P}_N \mathcal{G}^+(\mathcal{P}_N K; f, f) + \mathcal{I}_N \mathcal{G}^-(\mathcal{P}_N K; f, f), \quad f(t=0, \mathbf{v}) = \mathcal{P}_N F(t=0, \mathbf{v}), \quad (3.54)$$

$$\mathcal{I}_N \mathcal{G}^+(\mathcal{P}_N K; f, f) + \mathcal{I}_N \mathcal{G}^-(\mathcal{P}_N K; f, f), \quad f(t=0, \mathbf{v}) = \mathcal{P}_N F(t=0, \mathbf{v}), \quad (3.55)$$

$$\mathcal{I}_N \mathcal{G}^+(\mathcal{S}_N^\sigma K; f, f) + \mathcal{I}_N \mathcal{G}^-(\mathcal{S}_N^\sigma K; f, f), \quad f(t=0, \mathbf{v}) = \mathcal{P}_N F(t=0, \mathbf{v}), \quad (3.56)$$

$$\mathcal{I}_N \mathcal{G}^+(\mathcal{S}_N^\sigma K; f, f) + \mathcal{I}_N \mathcal{G}^-(\mathcal{S}_N^\sigma K; f, f), \quad f(t=0, \mathbf{v}) = \mathcal{I}_N F(t=0, \mathbf{v}). \quad (3.57)$$

Since  $K(\mathbf{p}, \mathbf{q})$  is a function in  $L^2_{\text{per}}([-T, T]^{2D})$ , here  $\mathcal{P}_N K$  means both projecting  $K(\mathbf{p}, \cdot)$  and  $K(\cdot, \mathbf{q})$  from  $L^2_{\text{per}}([-T, T]^D)$  to  $\mathbb{P}_N$ , and so is  $\mathcal{S}_N^\sigma K$ . The method (3.53) stands for the original FSM as described in Section 1.2. In the derivation, the kernel  $K$  is not explicitly projected. However, (1.17) shows that the result depends only on  $\mathcal{P}_N K$  since the Fourier series of  $f$  must be truncated when implementing the spectral method. It is not difficult to find that (3.54) is the first symmetric FSM introduced in Section 3.1.1, and (3.55) corresponds to the second symmetric FSM introduced in Section 3.2.2 (note that (3.55) is symmetric only for odd  $N$ ). Since a direct projection of  $K$  does not preserve the positivity of the kernel, the negative part of the discrete kernel may cause violation of the H-theorem. Nevertheless, all these three methods have spectral accuracy in the velocity space.

To ensure the positivity of the discrete kernel, the filter  $\mathcal{S}_N^\sigma$  is applied in (3.56), and thus positive coefficients (3.41) are obtained. The method (3.57) ensures the positivity of the approximation of  $f$  at collocation points, and thus the discrete H-theorem follows. However, the filter  $\mathcal{S}_N^\sigma$  has a significant smearing effect, which causes a reduction of the order of convergence. For any smooth periodic function  $f$ , the  $L^2$ -error  $\|f - \mathcal{S}_N^\sigma f\|_2$  is  $O(N^{-2})$ , and therefore the EMSM is at most second-order. The order of convergence will be numerically verified in the next section.

As a reference, we mention the positivity preserving method proposed in [20] for comparison purpose. The method uses the following discretization of the collision term:

$$\mathcal{S}_N^\sigma \mathcal{G}(\mathcal{P}_N K; f, f) + \mu \mathcal{S}_N^\sigma f - \mu f, \quad f(t=0, \mathbf{v}) = \mathcal{S}_N^\sigma F(t=0, \mathbf{v}), \quad (3.58)$$

where  $\mu$  is a constant dependent on  $\|F(t=0, \mathbf{v})\|_1$  such that  $\mu \geq \sup_{\mathbf{v} \in \mathbb{R}^D} \mathcal{L}(F)(t=0, \mathbf{v})$ .

## 4 Numerical tests

In this section, we perform several numerical tests for the space homogeneous Boltzmann equation to validate the accuracy, positivity and entropy monotonicity of the EMSM, and to compare with the positivity preserving spectral method in [20] and the classical FSM in [19]. First, we start by describing the implementation of the EMSM.

### 4.1 Implementation

In Section 2, we have presented the evolution equation of the EMSM (2.6), and pointed out the fast algorithms in [14, 7] are also valid for EMSM without destroying the H-theorem. In fact, one just needs to check that the fast algorithms do not destroy the first two conditions in Condition 1.

Given the definition of  $\hat{B}^\sigma(l, m)$  (3.47), in the fast algorithms in [14, 7],  $\hat{B}(l, m)$  are approximated by

$$\hat{B}(l, m) \approx \hat{B}_{fast}(l, m) := \sum_{p=1}^P \alpha_{l+m}^{(p)} \beta_l^{(p)} \gamma_m^{(p)}, \quad (4.1)$$

and  $\hat{B}_{fast}(l, m)$  satisfies the symmetry relation

$$\hat{B}_{fast}(l, m) = \hat{B}_{fast}(m, l) = \hat{B}_{fast}(-l, m). \quad (4.2)$$

Then  $\hat{B}^\sigma(l, m)$  can be approximated by

$$\hat{B}^\sigma(l, m) \approx \hat{B}_{fast}^\sigma(l, m) := \sum_{p=1}^P \alpha_{l+m}^{(p)} \left( \sigma_N(l) \beta_l^{(p)} \right) \left( \sigma_N(m) \gamma_m^{(p)} \right), \quad (4.3)$$



and  $\hat{B}_{fast}^\sigma(l, m)$  also satisfies the symmetry relation (4.2), which indicates the first condition in Condition 1 is valid for the fast algorithms. Next we check whether the fast algorithms destroy the non-negativity of  $\tilde{G}(\mathbf{p}', \mathbf{q}')$ .

To make it clear, we take the fast algorithm in [14] with  $D = 2$  and  $\tilde{B} = 1$  as an example. The first step of the fast algorithm in [14] is to write  $\mathbf{p}$  and  $\mathbf{q}$  in (1.18) in spherical coordinates  $\mathbf{p} = \rho \mathbf{e}$  and  $\mathbf{q} = \rho_* \mathbf{e}_*$  to get

$$\hat{B}(l, m) = \frac{1}{4} \int_{\mathbb{S}^1} \int_{\mathbb{S}^1} \delta(\mathbf{e} \cdot \mathbf{e}_*) \left[ \int_{-R}^R E_l(\rho \mathbf{e}) d\rho \right] \left[ \int_{-R}^R E_m(\rho_* \mathbf{e}_*) d\rho_* \right] d\mathbf{e} d\mathbf{e}_*. \quad (4.4)$$

Let  $\psi_R(l, \mathbf{e}) = \int_{-R}^R E_l(\rho \mathbf{e}) d\rho$ , then  $\hat{B}(l, m) = \frac{1}{4} \int_{\mathbb{S}^1} \int_{\mathbb{S}^1} \delta(\mathbf{e} \cdot \mathbf{e}_*) \psi_R(l, \mathbf{e}) \psi_R(m, \mathbf{e}_*) d\mathbf{e} d\mathbf{e}_*$ . Integrating it with respect to  $\mathbf{e}_*$  yields

$$\hat{B}(l, m) = \int_0^\pi \psi_R(l, \mathbf{e}_\theta) \psi_R(m, \mathbf{e}_{\theta+\pi/2}) d\theta. \quad (4.5)$$

The idea of the fast algorithm is to replace the integration in (4.5) with a quadrature formula. More precisely (4.5) is approximated by

$$\hat{B}_{fast}(l, m) = \sum_{p=1}^P \frac{\pi}{P} \psi_R(l, \mathbf{e}_{\theta_p}) \psi_R(m, \mathbf{e}_{\theta_p+\pi/2}). \quad (4.6)$$

Substituting (4.4) into (3.40) gives rise to

$$\tilde{G}(\mathbf{p}', \mathbf{q}') = \frac{1}{4} \int_{\mathbb{S}^1} \int_{\mathbb{S}^1} \delta(\mathbf{e} \cdot \mathbf{e}_*) \left[ \int_{-R}^R \chi_N^\sigma(\rho \mathbf{e} - \mathbf{p}') d\rho \right] \left[ \int_{-R}^R \chi_N^\sigma(\rho \mathbf{e}_* - \mathbf{q}') d\rho_* \right] d\mathbf{e} d\mathbf{e}_*. \quad (4.7)$$

Let  $\phi_R(\mathbf{p}', \mathbf{e}) = \int_{-R}^R \chi_N^\sigma(\rho \mathbf{e} - \mathbf{p}') d\rho$ . Apparently,  $\phi_R(\mathbf{p}', \mathbf{e}) \geq 0$  due to  $\chi_N^\sigma(\mathbf{p}) \geq 0$  for any  $\mathbf{p} \in \mathbb{R}^2$ . Then integrating (4.7) with respect to  $\mathbf{e}_*$  yields

$$\tilde{G}(\mathbf{p}', \mathbf{q}') = \int_0^\pi \phi_R(\mathbf{p}', \mathbf{e}_\theta) \phi_R(\mathbf{q}', \mathbf{e}_{\theta+\pi/2}) d\theta. \quad (4.8)$$

Similarly to (4.6), one obtains

$$\tilde{G}(\mathbf{p}', \mathbf{q}') \approx \sum_{p=1}^P \frac{\pi}{P} \phi_R(\mathbf{p}', \mathbf{e}_{\theta_p}) \phi_R(\mathbf{q}', \mathbf{e}_{\theta_p+\pi/2}). \quad (4.9)$$

Since  $\phi_R(\mathbf{p}', \mathbf{e}) \geq 0$  for any  $\mathbf{p}' \in \mathbb{R}^2$ ,  $\mathbf{e} \in \mathbb{S}^1$ ,  $\tilde{G}(\mathbf{p}', \mathbf{q}') \geq 0$  for any  $\mathbf{p}', \mathbf{q}' \in \mathbb{R}^2$ . Hence, the fast algorithm does not destroy the non-negativity of  $\tilde{G}(\mathbf{p}', \mathbf{q}')$ .

As we pointed out in Remark 2, the convolution rather than the anti-aliasing one can be directly used to calculate (2.6). Since the accuracy of EMSM is only second order, the smoothing filter is the main source of the error. In the fast algorithms, the number  $P$  perhaps can be smaller than that in [14, 7].

In the numerical simulation, the time can be discretized by kinds of discrete method, for example Runge-Kutta methods. In this paper, the third-order strong stability-preserving Runge-Kutta method proposed in [10] is employed in the discretization of time. And in all the tests, the time step is chosen as  $\Delta t = 0.01$ .

## 4.2 Numerical results

The test problems used here are solutions of the space-homogeneous Boltzmann equation for Maxwell molecules ( $\mathcal{B}(\mathbf{g}, \omega) = \frac{1}{2\pi}$  for  $D = 2$  and  $\mathcal{B}(\mathbf{g}, \omega) = \frac{1}{4\pi}$  for  $D = 3$ ).

**Example 1. 2D “BKW” solution** The first test problem is the well-known 2D “BKW solution”, obtained independently in [3] and [11]. The exact solution takes the form

$$F(t, \mathbf{v}) = \frac{1}{2\pi S} \exp\left(-\frac{|\mathbf{v}|^2}{2S}\right) \left(\frac{2S-1}{S} + \frac{1-S}{2S^2}|\mathbf{v}|^2\right), \quad (4.10)$$

where  $S = 1 - \exp(-t/8)/2$ . Using the “BKW” solution, we can check the accuracy, positivity of the solution and the entropy of the proposed method. Here we choose  $R = 6$  in the following numerical tests.

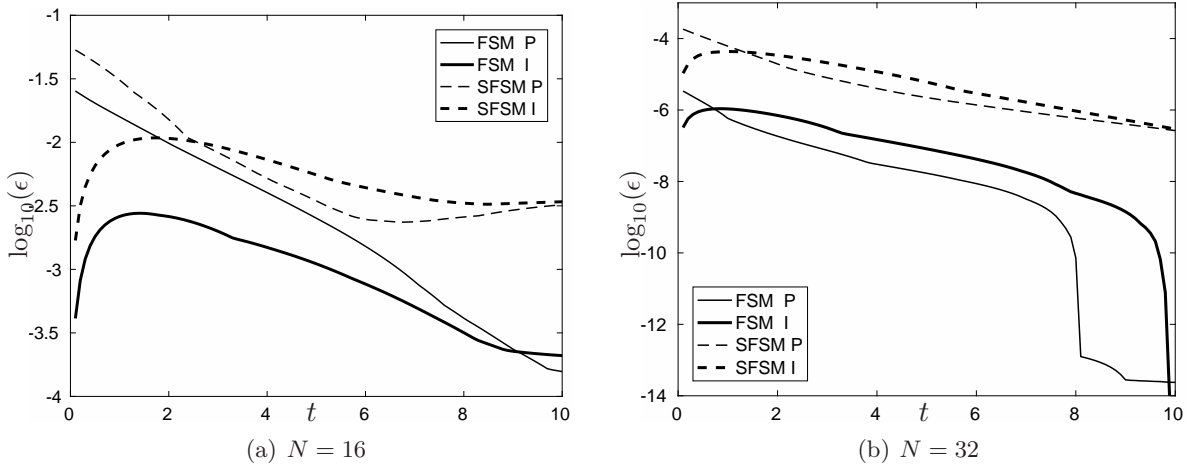


Figure 2: Positivity error of Fourier spectral method(FSM) (3.53) and symmetric Fourier spectral method(SFSM) (3.55) with the initial value given by orthogonal projection(P) and interpolation(I).

If we perform the computation with the FSM (3.53) or symmetry FSM (3.55), we can obtain good approximations of the exact solution at the collocation points. However, the solutions are not non-negative. The negative values of the solutions are partially caused by the initial value. However, even if we use interpolation rather than orthogonal projection to prepare the initial values, the solutions of these methods still fail to be non-negative. We measure the positivity error by

$$\epsilon := \frac{\sum_q |f(\mathbf{v}_q)| - \sum_q f(\mathbf{v}_q)}{\sum_q |f(\mathbf{v}_q)|}. \quad (4.11)$$

Figure 2 shows both methods fail to preserve the positivity of the solution at the collocation points no matter whether the initial value is given by orthogonal projection or interpolation. Thanks to the modification in (3.40), the positivity error of the EMSM (3.49) is zero.

Next, let us check the accuracy of (3.49). The  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$  errors and corresponding accuracy at  $t = 0.01$  is listed in the Table. 1. Here the  $\ell_p$  relative errors for  $p = 1, 2, \infty$

$N$	$\ell_1$ error	accuracy	$\ell_2$ error	accuracy	$\ell_\infty$ error	accuracy
16	$4.68 \times 10^{-3}$		$3.23 \times 10^{-3}$		$3.12 \times 10^{-3}$	
32	$1.72 \times 10^{-3}$	1.44	$1.36 \times 10^{-3}$	1.25	$1.40 \times 10^{-3}$	1.15
64	$5.54 \times 10^{-4}$	1.64	$4.56 \times 10^{-4}$	1.58	$5.57 \times 10^{-4}$	1.34
128	$1.55 \times 10^{-4}$	1.84	$1.29 \times 10^{-4}$	1.82	$1.73 \times 10^{-4}$	1.68
256	$4.05 \times 10^{-5}$	1.93	$3.42 \times 10^{-5}$	1.92	$4.73 \times 10^{-5}$	1.87

Table 1: The  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$  errors and corresponding accuracy for the BKW solution at  $t = 0.01$  with  $R = 6$ .

are defined by

$$\frac{\|f - F\|_p}{\|F\|_p} = \frac{\left(\sum_q |f(\mathbf{v}_q) - F(\mathbf{v}_q)|^p\right)^{1/p}}{\left(\sum_q |F(\mathbf{v}_q)|^p\right)^{1/p}}, \quad (4.12)$$

where the exact solution  $F$  are sampled in the calculation of the error estimate. Numerical results show the accuracy is second order, and the errors are also acceptable.

As discussed in Section 4.1, the fast algorithm in [14] can be applied to EMSM (3.49) to accelerate the computation. In (4.5), the integration on  $[0, \pi)$  can be reduced to  $[0, \pi/2)$ . Let  $M$  be the number of discrete points of  $[0, \pi/2)$ . Table 2 presents the  $\ell_1$  error for different values of  $M$ . One notices that  $M = 2$  is good enough in practice while in [14] the authors suggest  $M \geq 4$ .

$N$	$M = 2$	$M = 3$	$M = 32$
16	$4.6852 \times 10^{-3}$	$4.6826 \times 10^{-3}$	$4.6830 \times 10^{-3}$
32	$1.7241 \times 10^{-3}$	$1.7244 \times 10^{-3}$	$1.7245 \times 10^{-3}$
64	$5.5368 \times 10^{-4}$	$5.5388 \times 10^{-4}$	$5.5394 \times 10^{-4}$
128	$1.5485 \times 10^{-4}$	$1.5488 \times 10^{-4}$	$1.5489 \times 10^{-4}$
256	$4.0513 \times 10^{-5}$	$4.0516 \times 10^{-5}$	$4.0517 \times 10^{-5}$

Table 2:  $\ell_1$  error of the EMSM with fast algorithm in [14] for different number of discrete points of the circle.

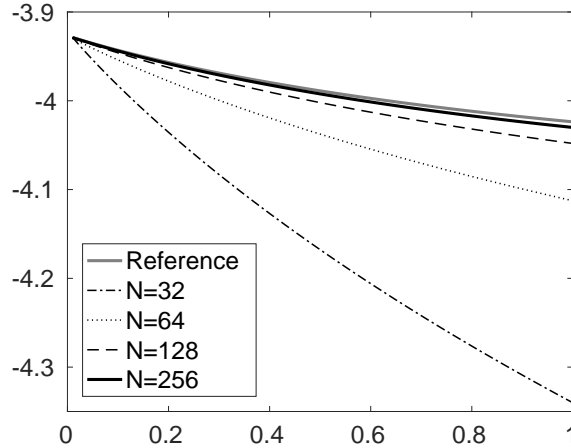


Figure 3: The evolution of the entropy of the EMSM (3.49) for different  $N$ .

The proposed method satisfies the H-theorem, so we study the evolution of the entropy. The discrete entropy is defined by

$$\eta = \left(\frac{2T}{N}\right)^D \sum_i f(\mathbf{v}_i) \log(f(\mathbf{v}_i)). \quad (4.13)$$

Then the evolution of the entropy is plotted in Figure 3. As the number of the discrete points  $N$  increases, the discrete entropy converges to the one of the exact solution.

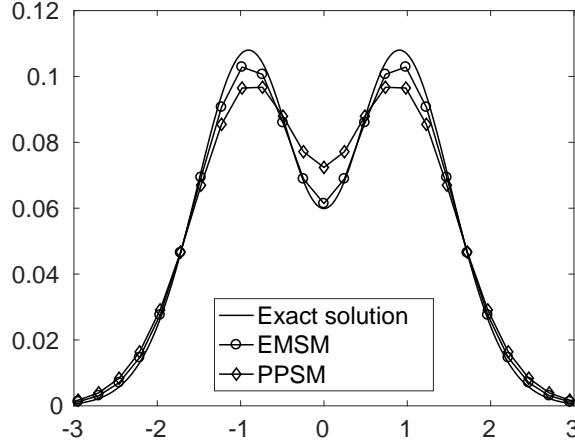


Figure 4: Comparison between the positivity preserving spectral method (PPSM) (3.58) proposed in [20] and the EMSM in (3.49) at  $t = 1$  with  $N = 64$  for BKW solution.

We also compare with the positivity preserving spectral method proposed in [20]. Figure 4 presents the numerical solutions in the  $\xi_1$  direction of the positivity preserving spectral method and the EMSM (3.49) at  $t = 1$  with  $N = 32$ . By contract, the smoothing filter used for the EMSM (3.49) results in less dissipation, thus leading to a better agreement with the exact solution. And as  $N$  increasing, the solution converges to the exact solution (see Figure 5).

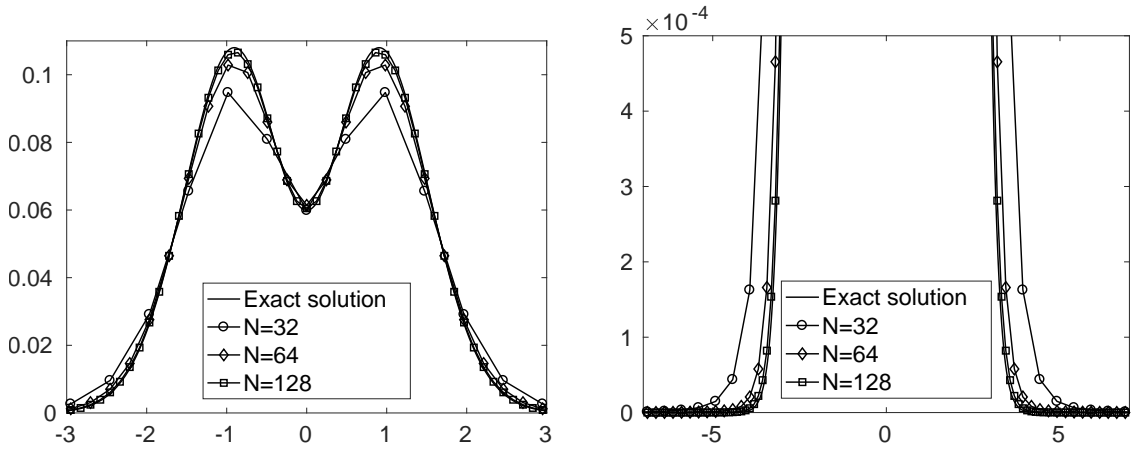


Figure 5: Numerical solution of the EMSM in (3.49) for different  $N$  with the BKW solution  $t = 1$  in different scales.

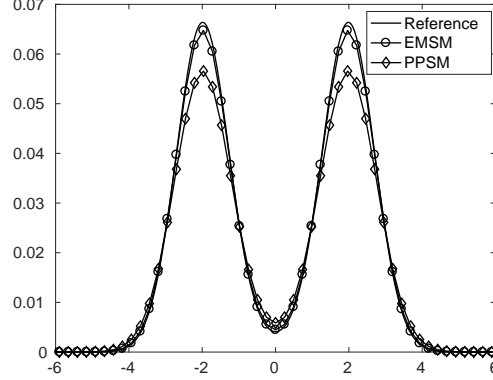


Figure 6: Comparison between the positivity preserving spectral method (PPSM) in [20] and the EMSM in (3.49) at  $t = 1$  with  $N = 64$  for the bi-Gaussian initial value.

**Example 2. Bi-Gaussian initial value** The second test, which is also frequently used, is a problem with the bi-Gaussian initial value equal

$$F(t=0, \mathbf{v}) = \frac{1}{4\pi} \left( \exp \left( -\frac{|\mathbf{v} - \mathbf{u}_1|^2}{2} \right) + \exp \left( -\frac{|\mathbf{v} - \mathbf{u}_2|^2}{2} \right) \right), \quad (4.14)$$

where  $\mathbf{u}_1 = (-2, 0)$  and  $\mathbf{u}_2 = (2, 0)$ . We solve it for the Maxwell molecules (2D in velocity), and set  $R = 6$ . Figure 6 shows the numerical results of the positivity preserving spectral method (PPSM) and the EMSM of this paper. The reference solution is calculated by the Fourier spectral method with  $N = 400$  and  $R = 8$ . The solution of the EMSM is close to the reference solution. And as  $N$  increases, it converges to the reference solution (see Figure 7).

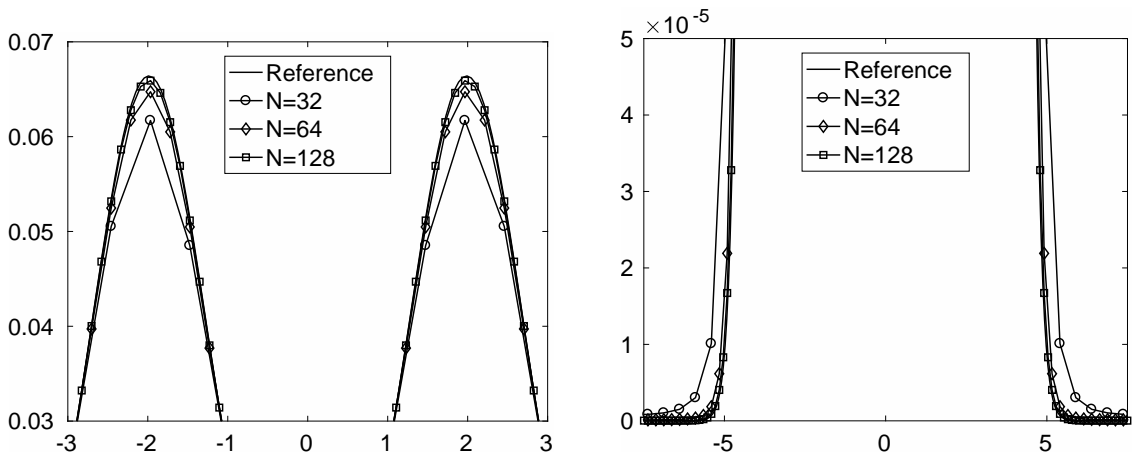
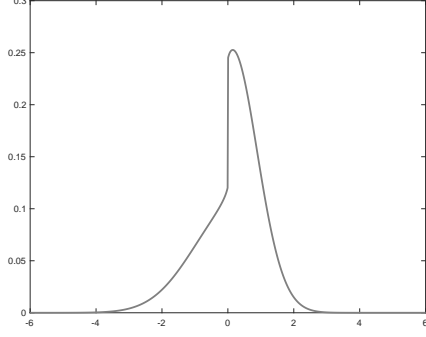
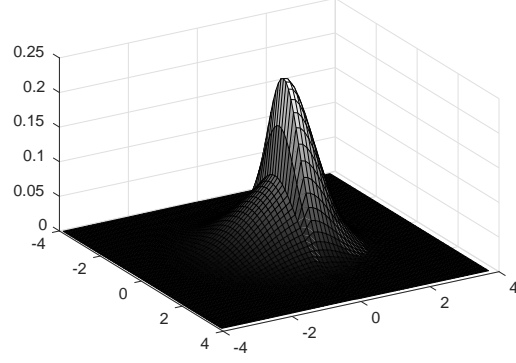


Figure 7: Numerical solution of the EMSM in (3.49) for different  $N$  with the bi-Gaussian initial value at  $t = 1$  in different scales.



(a) Profile of  $f(t = 0.5, \xi_1, \xi_2 = 0)$



(b) Profile of  $f(t = 0.5, v)$

Figure 8: Profile of  $f(t, v)$  with the discontinuous initial value (4.15) at  $t = 0.5$ .

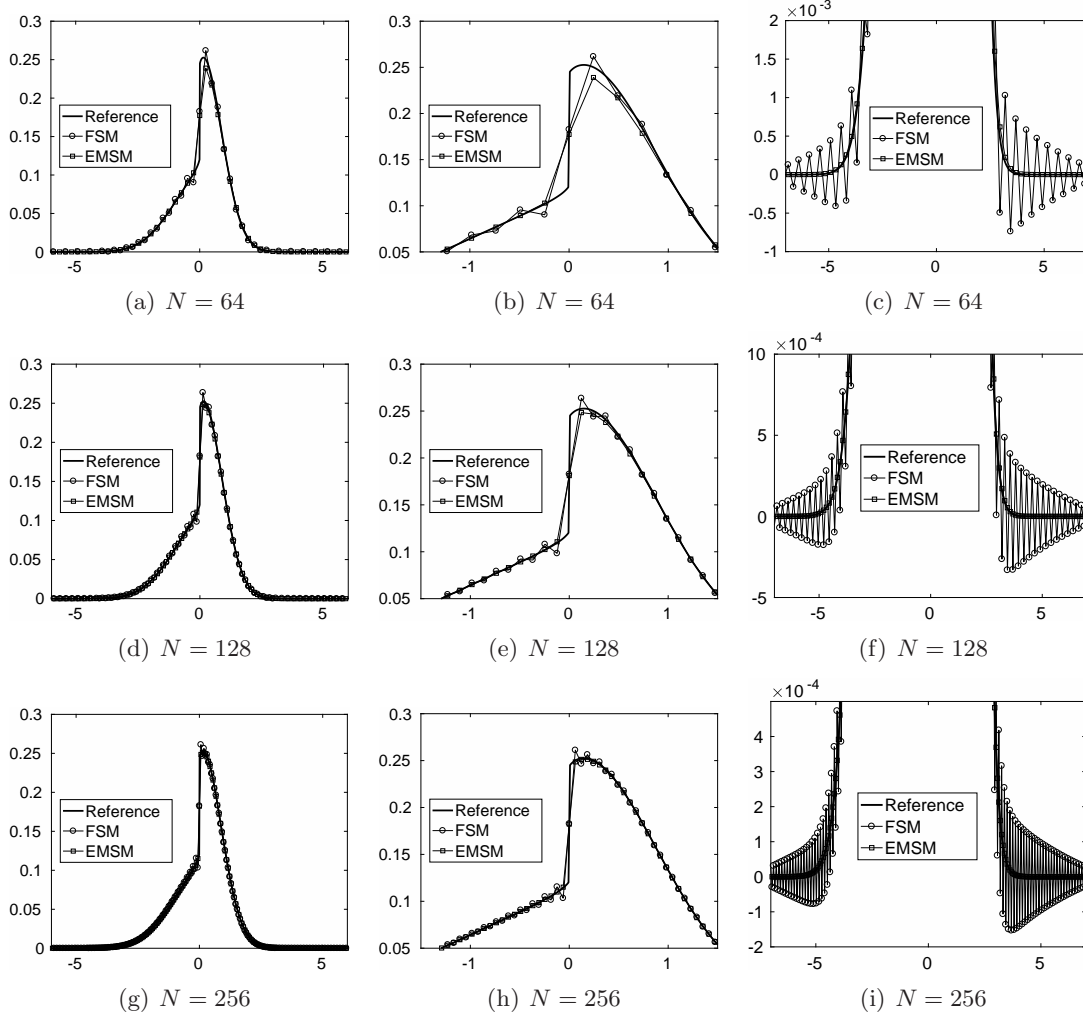


Figure 9: Numerical solution of the EMSM and FSM for different with the discontinuous initial value (4.15) at  $t = 0.5$  in different scale.

**Example 3. Discontinuous initial value** The third test problem comes with a discontinuous initial value

$$F(t=0, \mathbf{v}) = \begin{cases} \frac{\rho_1}{2\pi T_1} \exp\left(-\frac{|\mathbf{v}|^2}{2T_1}\right), & \text{if } \xi_1 > 0, \\ \frac{\rho_2}{2\pi T_2} \exp\left(-\frac{|\mathbf{v}|^2}{2T_2}\right), & \text{if } \xi_1 < 0. \end{cases} \quad (4.15)$$

In the test, we set  $\rho_1 = \frac{6}{5}$  and  $\rho_2, T_1$  and  $T_2$  are determined by

$$\int_{\mathbb{R}^D} f \, d\mathbf{v} = \int_{\mathbb{R}^D} f |\mathbf{v}|^2 / 2 \, d\mathbf{v} = 1, \quad \int_{\mathbb{R}^D} f \mathbf{v} \, d\mathbf{v} = 0.$$

The profile of the reference solution is presented in Figure 8, which is computed by the EMSM with  $N = 2048$ . Due to the discontinuous in the initial value, the spectral accuracy of FSM is lost. Instead the Gibbs phenomenon leads to oscillatory in the initial value. Numerical results in Figure 9 show around the discontinuity, the solutions of EMSM have a better agreement as compared to those of the FSM. The oscillation in the FSM solution cause a rather large error. The amplitude of the oscillatory decrease rather slowly as  $N$  increases. For the EMSM, there is no oscillatory around the discontinuous and the solution is smooth away from the discontinuous, i.e. no Gibbs phenomenon at all. This shows that the EMSM performs significantly better than the FSM for problems with discontinuous distribution functions.

**Example 4. 3D “BKW” solution** The fourth test problem is the 3D “BKW” solution, with the exact solution

$$F(t, \mathbf{v}) = \frac{1}{(2\pi S)^{3/2}} \exp\left(-\frac{|\mathbf{v}|^2}{2S}\right) \left(\frac{5S-3}{2S} + \frac{1-S}{2S^2} |\mathbf{v}|^2\right), \quad (4.16)$$

where  $S = 1 - 2\exp(-t/6)/5$ . Similarly to the 2D case, we first check the accuracy of the EMSM by this test problem. At  $t = 0.01$  with time step  $\Delta t = 0.01$ , the  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$  errors and corresponding accuracy is listed in the Table 3. Numerical results are similar

$N$	$\ell_1$ error	accuracy	$\ell_2$ error	accuracy	$\ell_\infty$ error	accuracy
16	$4.08 \times 10^{-3}$		$3.08 \times 10^{-3}$		$3.56 \times 10^{-3}$	
32	$1.42 \times 10^{-3}$	1.52	$1.12 \times 10^{-3}$	1.47	$1.26 \times 10^{-3}$	1.50
64	$4.07 \times 10^{-4}$	1.80	$3.29 \times 10^{-4}$	1.76	$3.72 \times 10^{-4}$	1.76
128	$1.08 \times 10^{-4}$	1.91	$8.85 \times 10^{-5}$	1.90	$1.00 \times 10^{-4}$	1.89

Table 3: The  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$  errors and corresponding accuracy for the BKW solution at  $t = 0.01$  with  $R = 6$ .

to the 2D case, i.e the accuracy is of the second order and the errors are rather small.

Next, we compare with the positivity preserving spectral method (PPSM) in [20]. Figure 10 presents the numerical solutions on the  $\xi_1$  direction of the positivity preserving spectral method and the EMSM (3.49) at  $t = 1$  with  $N = 32$ . It is clear that the smoothing filter used in the EMSM (3.49) results in much less dissipation, thus leading to a better agreement with the exact solution.

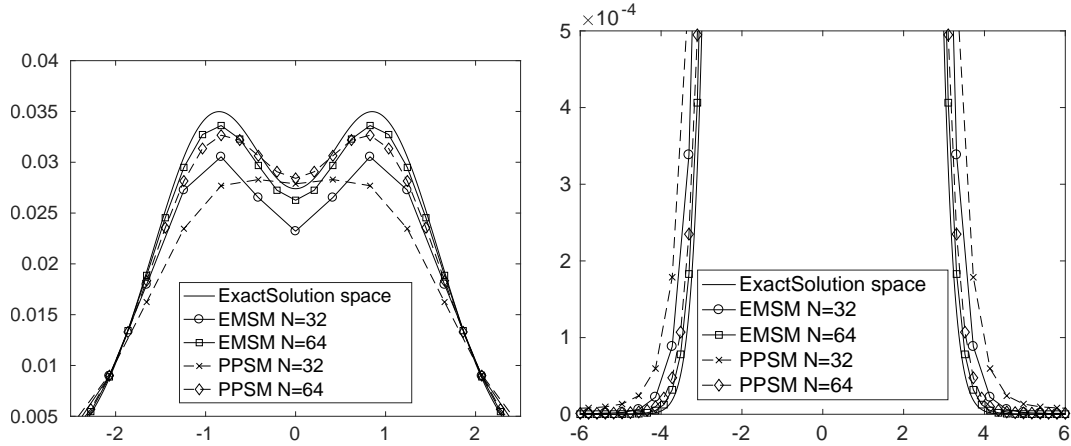


Figure 10: Numerical solution of the EMSM and PPSM for different  $N$  with the BKW solution  $t = 1$  in different scales.

## 5 Discussion

We would like to provide some comments on various methods based on uniform velocity grids. We start by listing some basic properties of major methods of this type. The classical DVM preserves a number of physical properties including positivity, the H-theorem, and the exact conservation of mass, momentum and energy but suffers from high computational cost and low accuracy. The classical FSM possesses spectral accuracy and an acceptable computational cost due to the fast algorithms in [14, 7], but several key physical properties are dropped to gain the high order accuracy. A fast algorithm for DVM is proposed in [15], and this method abandons the momentum and energy conservation.

The properties of these methods show that there seems to be a trade-off between the accuracy and the preservation of physical properties in the discretization of the collision term based on velocity grids. In the DVM, the conservation properties are well preserved due to a strict selection of collision pairs. One only allows collisions to happen if all the four velocities involved are exactly on the grid points, since this method does not utilize any interpolation or extrapolation to get distribution function values elsewhere. This is rather restrictive so that the approximation turns out to be very sparse and the accuracy drops below the first order. In the FSM, we allow all possible collisions by recovering the whole distribution function via interpolation, and thus a much better convergence rate is achieved. However, either projection or interpolation must be done after the evaluation of the collision term and this results in the loss of positivity and the H-theorem.

The entropy monotonic spectral method (EMSM) proposed in this paper is a trade-off between the accuracy and preservation of physical properties. It uses the ideas of both DVM and FSM in its derivation, and the scheme acts as a compromise: as for the convergence rate, the order is higher than DVM but lower than FSM; as for the physical properties, it keeps positivity, mass conservation and the H-theorem, while the momentum and energy conservation is lost; as for the computational cost, fast algorithms in [14, 7] are also valid for the EMSM.

Several other related methods exist. One of them is the positivity preserving spectral method (PPSM) in [20]. It also uses smoothing filters to guarantee the positivity of the solution. However, the PPSM introduces strong filters that bring significant numerical errors to the solution and the H-theorem still fails to hold. Our numerical results show



that the PPSM has larger numerical error than the EMSM introduced in this paper. Another related method, introduced in [13, 22], tries to improve the accuracy of DVM by interpolation. While the mass, momentum and energy are conserved in this scheme, positivity and the H-theorem fail to hold.

To sum up, we combine the ideas of the DVM and the FSM to obtain an entropy monotonic spectral method. Compared with the classical FSM, this method sacrifices the spectral accuracy to exchange the positivity of the solution at the collocation points and the H-theorem. When compared with the DVM, it sacrifices the conservation of momentum and energy in exchange for the second order accuracy and fast algorithm. We plan to study the loss of conservation of momentum and energy in the future work. Numerical simulations in the spatially inhomogeneous setting are in progress.

## References

- [1] G. A. Bird. Molecular Gas Dynamics and the Direct Simulation of Gas Flows. Oxford: Clarendon Press, 1994.
- [2] A Bobylev and S Rjasanow. Difference scheme for the boltzmann equation based on fast fourier transform. Technical report, 1996.
- [3] AV Bobylev. Exact solutions of the boltzmann equation. In Akademiia Nauk SSSR Doklady, volume 225, pages 1296–1299, 1975.
- [4] L. Boltzmann. Weitere studien über das wärmeleichgewicht unter gas-molekülen. Wiener Berichte, 66:275–370, 1872.
- [5] Torsten Carleman. Problemes mathématiques dans la théorie cinétique de gaz, volume 2. Almqvist & Wiksell, 1957.
- [6] Laura Fainsilber, Pär Kurlberg, and Bernt Wennberg. Lattice points on circles and discrete velocity models for the Boltzmann equation. SIAM Journal on Mathematical Analysis, 37(6):1903–1922, 2006.
- [7] Irene M Gamba, Jeffrey R Haack, Cory D Hauck, and Jingwei Hu. A fast spectral method for the Boltzmann collision operator with general collision kernels. pages 1–17, oct 2016.
- [8] Irene M Gamba and Sri Harsha Tharkabhushanam. Spectral-lagrangian methods for collisional models of non-equilibrium statistical states. Journal of Computational Physics, 228(6):2012–2036, 2009.
- [9] D Goldstein, B Sturtevant, and JE Broadwell. Investigations of the motion of discrete-velocity gases. Progress in Astronautics and Aeronautics, 117:100–117, 1989.
- [10] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. Strong stability-preserving high-order time discretization methods. SIAM review, 43(1):89–112, 2001.
- [11] Max Krook and Tai Tsun Wu. Exact solutions of the boltzmann equation. The Physics of Fluids, 20(10):1589–1595, 1977.
- [12] Günther Meinardus. Approximation of Functions: Theory and Numerical Methods. Springer, New York, 1967.

- [13] AB Morris, PL Varghese, and DB Goldstein. Improvement of a discrete velocity boltzmann equation solver that allows for arbitrary post-collision velocities. In Rarefied Gas Dynamics: Proceedings of the 26th International Symposium, Kyoto, Japan, 2008.
- [14] C. Mouhot and L. Pareschi. Fast algorithms for computing the Boltzmann collision operator. Math. Comp., 75(256):1833–1852, 2006.
- [15] Clément Mouhot, Lorenzo Pareschi, and Thomas Rey. Convolutional decomposition and fast summation methods for discrete-velocity approximations of the Boltzmann equation. ESAIM: Mathematical Modelling and Numerical Analysis, 47(5):1515–1531, 2013.
- [16] Andrzej Palczewski, Jacques Schneider, and Alexandre V. Bobylev. A Consistency Result for a Discrete-Velocity Model of the Boltzmann Equation. SIAM Journal on Numerical Analysis, 34(5):1865–1883, oct 1997.
- [17] Vladislav A. Panferov and Alexei G. Heintz. A new consistent discrete-velocity model for the Boltzmann equation. Mathematical Methods in the Applied Sciences, 25(7):571–593, 2002.
- [18] Lorenzo Pareschi and Benoit Perthame. A fourier spectral method for homogeneous boltzmann equations. Transport Theory and Statistical Physics, 25(3-5):369–382, 1996.
- [19] Lorenzo Pareschi and Giovanni Russo. Numerical solution of the boltzmann equation i: Spectrally accurate approximation of the collision operator. SIAM journal on numerical analysis, 37(4):1217–1245, 2000.
- [20] Lorenzo Pareschi and Giovanni Russo. On the stability of spectral methods for the homogeneous boltzmann equation. Transport Theory and Statistical Physics, 29(3-5):431–447, 2000.
- [21] François Rogier and Jacques Schneider. A direct method for solving the boltzmann equation. Transport Theory and Statistical Physics, 23(1-3):313–338, 1994.
- [22] PL Varghese. Arbitrary post-collision velocities in a discrete velocity scheme for the boltzmann equation. In Proc. of the 25th Intern. Symposium on Rarefied Gas Dynamics/Ed. by MS Ivanov and AK Rebrov. Novosibirsk, pages 225–232, 2007.
- [23] Alexandre Vasiljevitch Bobylev, Andrzej Palczewski, and Jacques Schneider. On approximation of the boltzmann equation by discrete velocity models. Comptes rendus de l’Académie des sciences. Série 1, Mathématique, 320(5):639–644, 1995.
- [24] Alexander Weiße, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. The kernel polynomial method. Reviews of modern physics, 78(1):275, 2006.